ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение высшего профессионального образования
«Уральский государственный университет им. А.М. Горького»

ИОНЦ «Информационная безопасность»

математико-механический факультет

кафедра алгебры и дискретной математики

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС
«Морфологические анализаторы русского языка»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ИСПОЛЬЗОВАНИЮ КОМПЛЕКСА

Автор: ведущий математик РУНЦ
«Информационная безопасность»
Ю. С. Лукач

Екатеринбург
2008

ВВЕДЕНИЕ

Учебно-методический комплекс «Морфологические анализаторы русского языка», далее именуемый УМК, представляет собой библиотеку классов для платформы Microsoft .NET 2.0, позволяющую производить детальный морфологический анализ русского текста. Оформление УМК в виде библиотеки позволяет студентам и преподаваиелям с легкостью использовать его для решения различных задач, связанных с анализом естественного текста. Тем самым, данный УМК может быть полезен при изучении дисциплин, связанных с синтаксическим и семантическим анализом языковых конструкций, в частности, математической лингвистики, лингвистических основ информатики, теории формальных грамматик и компиляции.

УМК имеет следующие отличительные особенности.

- В него включена детальная морфологическая информация более чем о 205 тыс. русских слов (около 4 млн. словоформ). Вся эта информация получена путем автоматического порождения всех возможных словоформ русского языка из лексической базы данных, составленной автором путем сведения воедино всех авторитетных словарей русского языка.
- Морфологическая информация предельно компактифицирована и занимает на диске всего около 15 Мб (при исходном размере текстовых файлов более 200 Мб). В начале работы с библиотекой вся эта информация считывается в оперативную память коспьютера для обеспечения максимального быстродействия.
- Поиск базовой формы и грамматических дескрипторов для данной словоформы производится с помощью минимального конечного распознавателя, разработанного автором. Это гарантирует нахождение информации о словоформе, состоящей из *n* букв, за *n* шагов распознавателя, обеспечивая тем самым максимально возможную скорость морфологического анализа текста.
- В зависимости от потребностей пользователя возможен как минимальный (с получением только базовой формы слова), так и предельно полный морфологический анализ (с получением всех возможных базовых форм, их лексических и грамматических характеристик, а также описания всех способов порождения данной словоформы из каждой базовой формы).

СОСТАВ УМК

В состав УМК входят библиотека Morpho версии 2.1, ее описание и демонстрационная программа.

С точки зрения пользователя Morpho представляет собой обычную библиотеку классов Microsoft .NET 2.0, которая предоставляет возможности, перечисленные в предыдущем разделе.

Дистрибутивный диск имеет следующую структуру папок:

- BIN –библиотека Morpho.dll, исполняемый демонстрационный файл MorphoDemo.exe и файлы морфологических данных Common.bin, Common.dic и Common.din.
- SRC – исходные тексты программ Morpho и MorphoDemo, а также файлы проектов для Microsoft Visual Studio 2005.
- В корневой папке диска находятся файлы документации: титульный лист УМК (Title.pdf), данный файл (Readme.pdf) и руководство пользователя (Guide.pdf).

Для освоения библиотеки Morpho следует ознакомиться с руководством пользователя и исходными текстами демонстрационной программы.

Замечания, сообщения об ошибках и предложения по улучшению функциональности библиотеки прошу направлять разработчику по адресу yury@usaaa.ru.